Mathematical tools for automatic program analysis

Philippe Flajolet

INRIA

UNITÉ DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél. (1) 39 63 55 11

Rapports de Recherche

N° 603

MATHEMATICAL TOOLS FOR AUTOMATIC PROGRAM ANALYSIS

Philippe FLAJOLET

Février 1987

Mathematical Tools for Automatic Program Analysis

Outils mathèmatiques pour l'analyse automatique de programmes

Philippe Flajolet

Abstract: *This report consists of two notes by the author. Both are informal presentations of mathematical methods that may be employed in order to automatically extract informations on the average case complexity of some well-characterized classes of algorithms and programs. Such methods are amenable to implementation using symbolic manipulation systems.*

Résumé: Ce rapport consiste en deux notes qui présentent informellement les méthodes mathématiques que l'on peut utiliser pour extraire automatiquement des informations sur la complexité en moyenne de classes bien définies d'algorithmes et de programmes. Ces méthodes sont implémentables grâce à l'utilisation de systèmes de calcul formel.

---

1. *Mathematical and Computational Tools for an "Assistant Programme Analyser"*, November 1986. This is a short abstract submitted to the EUROCAL'87 Conference (Leipzig, June 1987).
2. *Elements of a General Theory of Combinatorial Structures*. This is a version of an invited lecture given at the FCT'85 Conference in Cotbus, GDR, in September 1985. The paper appears in:
   Fundamentals of Computation Theory, Lecture Notes in Computer Science **199** (1985), pp. 112-127.

Mathematical and Computational Tools for an "Assistant Programme Analyser"

Philippe Flajolet
INRIA, 78150 Rocquencourt (France)

The field of average-case analysis of discrete algorithms and data structures has been largely founded by Knuth (see *The Art Of Computer Programming*, esp. vol. III). Given an algorithm for sorting or searching, like a binary search tree or heap-ordered tree algorithm, a hashing method of sorts *etc.*, the problem is to determine its expected behaviour under some reasonable random input model. The *analysis* of an algorithm decomposes into an exact combinatorial counting problem, followed by asymptotic estimations meant to express the average complexity in terms of a standard asymptotic scale, comprising such functions as n^3, $n^{2.81}$, $n \log n$...

Our past experience in the analysis of algorithms seems to indicate that a relatively restricted number of "schemes" cover a majority (not all!) of existing analyses. To mention a specific example [1], the *digital search tree* model appeared independently in: (i) the analysis of digital searching with *tries*, (ii) dynamic and extendible hashing algorithms for files on disk, (iii) network communication protocols, (iv) probabilistic estimation algorithms in data bases, (v) some polynomial factorization algorithms.

Such schemes are indeed methods that permit the exact and/or asymptotic analysis of a large (and well-characterised) class of parameters over certain combinatorial structures. The systematic nature of those methods makes it tempting to transform them into algorithms and programmes that will take as input a specification of some combinatorial counting problem and deliver, as output, asymptotic estimations.

We feel that such a task can be initiated using "off the shelf" technology in combinatorial analysis and complex asymptotics. Our purpose here to indicate succintly what this technology is and how it can be used to achieve our goal, putting in perspective some recent or on-going research [1,2,3,4,5,6].

1. Compiling Combinatorial Descriptions into Generating Functions Equations:

Around 1910-1920, Major Percy MacMahon developed a rather algebraic view of exact enumerations via generating functions. That approach was revived in the 1960's by Rota, Foata and Schützenberger, and later Joyal, then applied in the area of the analysis of algorithms by the author [2] and D. Greene (PhD Thesis, Stanford Univ., 1984). It is exposed, in a systematic way, in the recent treatise by Jackson and Goulden (*Combinatorial Enumerations*, 1983).

The basic idea of the approach – called the *symbolic method*– is that a fairly large collection of *set-theoretic constructions* translate into *functionals* over generating functions.[§] Thus, for instance, if $C = A \cup B$. $C = A \times B$, $C = A^*$ (the sequence-of construction), $C = 2^A$ (power-set construction), the corresponding ordinary generating functions satisfy:

$$c(z) = a(z) + b(z)\,;\ c(z) = a(z).b(z)\,;\ c(z) = \frac{1}{1-a(z)}\,;\ c(z) = \exp\left(a(z) + \Phi[a(z)]\right).$$

where $\Phi[.]$ is the Pólya operator:

$$\Phi[f(z)] = -\frac{f(z^2)}{2} + \frac{f(z^3)}{3} - \frac{f(z^4)}{4} + \cdots$$

Extensions to labelled structures exist based on Foata's partitional concept (or almost equivalently Greene's labelled grammars).

The class of *elementary (iterative) structures* is defined as the closure of the finite sets under the above operations. The class of *elementary recursive structure* is defined as a similar closure when recursive definitions are allowed.

For elementary iterative structures, a simple tree transducer will translate a structural definition into an explicit form for the associated generating function. For elementary recursive structures, a set of fixed point functional equations will be obtained. We have:

Proposition 1: *It is decidable whether an elementary structure definition is well founded. For a well founded definition, the corresponding counting problem can be solved in a polynomial number of operations.*

[§] If C is a class of structures with c_n the number of structures of size n, the ordinary generating function (ogf) of C is $c(z) = \sum c_n z^n$ and the exponential generating function (egf) is $\sum c_n \frac{z^n}{n!}$

One first needs to check that the operator associated to the specification is contracting. This may be done by a reduction to the empty word problem for context-free languages. Next observe that the method of "indeterminate coefficients" can be systematically applied in polynomial time. ■

M. Soria (Orsay) has implemented that method (in the language ML). Her programme compiles such structural specifications into functions/equations in the MAPLE symbolic manipulation system, and effectively produces the corresponding (numerical) counting results. Greene (*op. cit.*) had an earlier implementation for a special subset of the above constructs, based on recurrences. Such systems pose, in passing, a number of interesting optimization problems.

Examples of elementary structures include:

> Regular languages; unambiguous context-free languages; integer partitions; integer compositions; permutations with local order constraints (like up-and-down permutations); finite functions; occupancy distributions; trees (labelled and unlabelled) *etc.*

What is also of interest is that, along with elementary structures, one can define a notion of elementary parameters. Statistics for those parameters also satisfy a form of Prop. 1. Typical examples are: number of summands for partitions and compositions; path length for trees; cycle length for finite functions *etc.* Those have obvious applications to the analysis of algorithms since simple algorithms have costs usually expressible in terms of such elementary parameters.

2. Analytic Functionals and Asymptotic Counting:

The functionals that apppeared in the previous section are not "random": we could call them *analytic* (they map functions analytic at the origin into themselves) *monotone functionals* since they preserve majorance relations between functions: if $f \ll g$ in the sense that $|f_n| < |g_n|$, then we also have $\Psi(f) \ll \Psi(g)$.

By the Darboux-Pólya class of methods (see Henrici's book or [3] for expository material), it is known that, for functions with isolated singularities, the asymptotic behaviour of the function near its dominant singularities (the ones of smallest modulus) determines the asymptotic behaviour of the Taylor coeficients of the function. Furthermore, by Pringsheim's theorem, all our generating functions should be expected to have a dominant positive real singularity.

These simple observations have important consequences for asymptotics since they lead to partial decision procedures regarding asymptotic estimates that cover many cases of interest in applications.

Assume for instance that $C = A^*$: class C is formed of all sequences $(u_1, u_2, \ldots, u_k)$ where $u_j \in A$ and k is an integer. If the ordinary generating function (ogf) $a(z)$ of A is aperiodic ($a(z) \neq \alpha(z^p)$ for any $p \geq 2$) and is infinite at its dominant real singularity, then the ogf $c(z) = (1 - a(z))^{-1}$ will have a unique dominant singularity that is a real pole. If we set $\rho = a^{(-1)}(1)$, then:

$$c_n \sim \frac{1}{\rho a'(\rho)} \rho^{-n}.$$

Thus the asymptotic counting of class C can be done with very little information (also effectively computable to any degree of accuracy) on $a(z)$. To take a particular example, let A be the set of (non-zero) integer partitions, so that $C = A^*$ is the collection of *sequences* of partitions, and c_n is the number of such sequences with total weight n. It is well known that:

$$1 + a(z) = \prod_{j=1}^{\infty} \frac{1}{1 - z^j}.$$

The asymptotic analysis of a_n is difficult (it was done by Hardy and Ramanujan, and in its full generality requires some theory of elliptic functions). However, any reasonable symbolic manipulation system will enable us to find easily that:

$$c_n \simeq 0.4141137931 \times 2.6983291064^n$$

Fact 2: *For elementary iterative structures, there is a partial algorithmic procedure to determine the asymptotics of their counting function.*

The principle is as follows: Walking up the specification tree, one can (partly) compute inductively attributes like: "is-a-polynomial", "is-entire", "is-aperiodic", "is-infinite-at-singularity". From there, observing that the radii of convergergence decrease as we walk up the tree, one can determine a *critical* functional that determines the value of the dominant singularity of the generating function. Functionals on top of this critical functional then determine the *nature* of that singularity. ■

The above theorem is actually the basis of a typology of iterative structures (see [4] for a detailed example concerning "meromorphic" structures). In the case of recursive structures, extending results from Meir and Moon (*Canad. J. of Math*, 1978), we get [5]:

Theorem 3: *For elementary recursive structures defined using only unions, cartesian products and sequence-of constructs, the asymptotic counting problem is effectively computable.*

Corresponding generating functions are algebraic. Using elimination followed by applications of Newton's polygon rule, their singular behaviours can be effectively determined. Conclude via a proper use of the Darboux-Pólya method. ■

Only partial results are available so far when a richer set of constructions is used. Notice however that, in many cases, the appearance of the Φ operator inside recursive definitions need not be dramatic since for any ogf $f(z)$ with radius of convergence < 1 $\Phi[f(z)]$ has a radius of convergence strictly larger than that of $f(z)$. (See the counting of non-planar unlabelled trees in graph enumerations and some of Pólya and Otter's original examples).

3. Extensions:

Our original problem of analyzing algorithms leads not only to the counting of structures but also to estimating asymptotically expectations and/or distributions of various structural parameters. Some results have been earlier obtained by Bender, but they only deal with the application of a single functional (labelled sequence-of or set-of) to an explicitly given structure. In [4], we indicate principles upon which a typology of structural parameters can further be built. As only illustration, we shall cite:

Theorem 4: *Let A, B and C be labelled structures such that:*

$$C=\text{sequence-of}(B) \text{ with } B=\text{set-of}(A)$$

and assume that the sequence-of *construction is critical in C. Then, the proportion of B-component in a random C structure containing exactly k A-objects tends a limiting Poisson distribution with finite mean.*

A number of similar results (with geometric or Gaussian distributions) regarding other general combinatorial schemas can be obtained. Of course, all corresponding parameters can be computed explicitly (and conveniently, using symbolic manipulation systems).

Again, for recursive structures, our knowledge is less advanced. However in [6], we describe a simple programming language on trees – PL-TREE – for which systematic translation mechanisms exist (to generating functions and asymptotics). That language is powerful enough to express non trivial pattern matching algorithms. There also, phenomena comparable to Theorem 4 can be detected.

Conclusions:

From what we often read in the literature, artificial intelligence systems distinguish themselves by powerful reasoning capabilities. The system we propose uses instead a rather poor and limited form of logic, and is based on a modest implementation of results from 19th century style mathematics. However, in the course of our investigations, it appeared that the combined study of algebraic and analytic functionals arising in combinatorial enumerations leads to results that are not only simple and somewhat powerful but also amenable to implementation using current symbolic manipulation systems like MACSYMA or MAPLE.

References:

[1] Algebraic Methods for Trie Statistics, *Annals Discr. Math.* **25** (1985), pp. 145-188. (With M. Regnier, D. Sotteau)

[2] Analyse d'algorithmes de manipulation d'arbres et de fichiers, *Cahiers du BURO* **31** (1981), 209p.

[3] Lecture Notes on the Analysis of Algorithms, Princeton University (1986).

[4] Elements of a General Theory of Combinatorial Structures, *Lect. Notes in Comp.Sc.* **199** (1985), pp 112-127.

[5] Analytic Models and Inherent Ambiguity of Context Free Languages, *Theoret. Comp. Sc.* (1987), to appear. (Preliminary version in *Lect. Notes in Comp. Sc.*, **194**.)

[6] A Complexity Calculus for Recursive Search Programs over Tree Structures, *Math. Syst. Th.* (1987), to appear. Preliminary version in Proc. IEEE FOCS 1982. (With J-M. Steyaert.)

ELEMENTS OF A GENERAL THEORY OF COMBINATORIAL STRUCTURES

Philippe Flajolet

I.N.R.I.A.
Rocquencourt
F-78150 Le Chesnay (France)

ABSTRACT

This paper presents some preliminary observations relating in many cases structural definitions of combinatorial structures to statistical properties of their characteristic parameters.
The developments are based on two observations: (i) for a large family of classes of combinatorial structures, one can compile *structural descriptions into functional equations over counting generating functions; (ii)* general analytical patterns *arise from the study of these functional equations.*
As a consequence, statistical evaluations of a large number of parameters of combinatorial structures can be automated using symbolic manipulation systems.
The approach taken also suggests the existence of general theorems concerning statistical properties of combinatorial structures that may be used to analyse combinatorial structures of a complex form.

1. Introduction.

A *class of combinatorial structures* in the widest sense is a finite or denumerable set **C** together with a *size* function denoted $|.|$ or $|.|_{\mathbf{C}}$ such that for all integer n, the set of objects in **C** of size n is finite. The *counting problem* for **C** consists in determining the sequence $\{c_n\}_{n\geq 0}$ defined by

$$c_n \equiv card\{\omega \in \mathbf{C} \mid |\omega| = n\}$$

Classes usually considered in combinatorial analysis are formed of permutations, words, trees, graphs, finite functions *etc*... . The size of a permutation or a word is the number of elements (letters) it comprises; the size of a graph or a tree is its number of nodes ... .

A *combinatorial construction* is an operation, in the normal set-theoretic sense, on classes of combinatorial structures. Its specification includes a description on how the size of the result of the construction can be obtained from the sizes of the operands.

As an example, the *Cartesian product* construction associates to each pair **C**, **D** of classes a class **E** defined as usual by:

$$\mathbf{E} = \mathbf{C}\times\mathbf{D} = \{(\gamma,\delta) \mid \gamma \in \mathbf{C}\,,\, \delta \in \mathbf{D}\}$$

with the natural notion of size:

$$|(\gamma,\delta)|_{\mathbf{E}} = |\gamma|_{\mathbf{C}} + |\delta|_{\mathbf{D}} .$$

The Algebra of Generating Functions. The most convenient way of approaching counting problems is via the use of generating functions. The *ordinary generating function* (*o.g.f.*) of class **C** is:

$$c(z) = \sum_{n \geq 0} c_n z^n .$$

The *exponential generating function* (*e.g.f.*) of class **C** is:

$$\hat{c}(z) = \sum_{c_n} \frac{z^n}{n!} .$$

A combinatorial construction K is *admissible* if the counting sequence of the result, say $\mathbf{E} = K(\mathbf{C},\mathbf{D})$ when K is binary, is only determined by the counting sequences of the arguments (here **C** and **D**). In that case, there exist *operators* Φ and $\hat{\Phi}$ over formal power series such that:

$$e(z) = \Phi(c(z),d(z)) \; ; \; \hat{e}(z) = \hat{\Phi}(\hat{c}(z),\hat{d}(z)) .$$

(Note: here and in the sequel, we adhere to the notational convention of representing systematically classes of combinatorial structures, counting sequences and corresponding generating functions by the same groups of letters).

Returning to the example of the cartesian-product construction, we see by a direct argument that:

$$e_n = \sum_{k=0}^{n} c_k d_{n-k}$$

so that the cartesian-product construction is admissible. The corresponding operator over ordinary generating functions is:

$$e(z) = c(z) . d(z) .$$

A collection of admissible constructions is given for instance in [GJ83] or [Fla85a]. In particular, [GJ83] show how most of classical combinatorial analysis can be nicely expressed using that framework. We shall use some of the following admissible constructions (listed together with the corresponding operators on either ordinary or exponential generating functions):

A. *Ordinary generating functions:*

(disjoint) union	$\mathbf{E}=\mathbf{C}\cup\mathbf{D}$	=>	$e(z)=c(z)+d(z)$
cartesian product	$\mathbf{E}=\mathbf{C}\times\mathbf{D}$	=>	$e(z)=c(z).d(z)$
sequence-of	$\mathbf{E}=\mathbf{C}^*$	=>	$e(z)+(1-c(z))^{-1}$
set-of	$\mathbf{E}=2^{\mathbf{C}}$	=>	$e(z)=\exp(c(z)-\frac{1}{2}c(z^2)+\frac{1}{3}c(z^3)\cdots)$
substitution	$\mathbf{E}=\mathbf{C}[\mathbf{D}]$	=>	$e(z)=c(d(z))$

(In all these constructions the size of the result is the sum of the sizes of components)

B. *Exponential generating functions:*

(disjoint) union	$\mathbf{E}=\mathbf{C}\cup\mathbf{D}$	=>	$\hat{e}(z)=\hat{c}(z)+\hat{d}(z)$

partitional product (P-product)	$E=C^{\bullet}D$	=>	$\hat{e}(z)=\hat{c}(z).\hat{d}(z)$
partitional complex (P-sequence-of)	$E=C^{<\bullet>}$	=>	$\hat{e}(z)=(1-\hat{c}(z))^{-1}$
abelian partitional complex (P-set-of)	$E=C^{[\bullet]}$	=>	$\hat{e}(z)=exp(\hat{c}(z))$

(This last set of constructions formally introduced by Foata operate like the classical constructions but they deal with *well-labelled objects* and distribute labels in all possible ways consistent with the labellings of the arguments).

Thus many combinatorial constructions translate into operators over generating functions. *Direct* constructions of a class from trivial (*e.g.* finite) sets will lead to *explicit* equations for counting generating functions; *indirect* constructions (*e.g.* recursive) will lead to *functional equations* determining the counting generating functions *implicitly*.

The Analysis of Generating Functions: Another important component of statistics for parameters of combinatorial structures comes from consideration of analytic properties of generating functions.

To start with a simple example, if $f(z)$ is a (counting) generating function, and if we may determine that its radius of convergence is ρ, then for arbitrary ε, one has for the coefficients f_n the inequalities:

$$\rho^{-n}(1-\varepsilon)^n <_{i.o.} f_n <_{a.e.} \rho^{-n}(1+\varepsilon)^n$$

where "$<_{i.o.}$" means "less than infinitely often" (for infinitely many values of n) and "$<_{a.e.}$" means "less than almost everywhere" (except for possibly a finite set of values of n).

Thus the radius of convergence (*r.o.c.*) of the counting generating function (which is also determined by the modulus of the singularity nearest to the origin) carries useful information on the growth rate of the counting sequence.

More refined estimates are usually available. The fundamental tool is Cauchy's integral formula:

$$f_n = \frac{1}{2i\pi}\int_{\Gamma} f(z)\,\frac{dz}{z^{n+1}}\,.$$

By taking contours that come close to the dominant singularities, using *Darboux' method* [He77] or the type of *singularity analysis* of [FO83], or by using *saddle point* methods [He77],[DB57], one is often able to determine the asymptotic form of coefficients of functions either defined explicitly or accessible through some functional equations.

As an example, the exponential generating function for the class **F** of permutations without cycles of length 1 or 2 is:

$$\hat{f}(z) = \frac{e^{-z-z^2/2}}{1-z}$$

and from the singular expansion (valid near the singularity $z=1$):

$$\hat{f}(z) \sim \frac{e^{-3/2}}{1-z}$$

one is able to deduce:

$$\frac{f_n}{n!} \sim [z^n]\,\frac{e^{-3/2}}{1-z} \sim e^{-3/2}$$

where we have used the classical notation $[z^n]f(z)$ to represent the coefficient of z^n in the Taylor expansion of $f(z)$.

We are now in a position to explain our main goal in this paper:

Consider the functionals (operators) associated to admissible combinatorial constructions. These functionals can also be viewed as analytic functionals, *that is to say functionals mapping analytic functions into analytic functions. Determine to what extent analytic properties of these functionals should reflect statistical properties of of combinatorial constructions of which they are images.*

2. Classical Examples.

From the preceding discussion, we are interested in classes of combinatorial structures defined from trivial classes by closure under a set of admissible constructions $\{K_1, K_2, \cdots, K_m\}$ that have the further property that the chain:

structural definitions → *functional equations* → *asymptotic analysis*

can be followed automatically. This requires a characterisation of some form of the corresponding counting generating functions from which (via singularity analysis or saddle point methods say) general asymptotic results will derive.

To illustrate our subject, we briefly recall 3 known cases: the family of regular languages and the family of context-free languages in formal language theory ; the family of simple classes of trees in combinatorics.

Regular Languages and Regular Events:

A language is a set of words over a finite alphabet. The family of regular languages is defined as the closure of the family of finite sets (of words) under the operations of union, catenation product and star operation [Ei74]. It more or less corresponds to the family of combinatorial structures obtained from the finite classes by means of union, cartesian product and the sequence-of construct.

From the translation mechanism recalled in the introduction, there follows that regular languages have generating functions obtained from polynomials by sequences of sum, product, and quasi-inverse operator $Q(y)=(1-y)^{-1}$. (To be precise, this requires some consideration about making definitions deterministic.) Thus the generating functions of regular languages are *rational*. Now the asymptotic behaviour of coefficients of rational function is well known: such a coefficient is expressible asymptotically as a combination of terms of the form:

$$C\alpha^n n^r \tag{1}$$

where C, α are algebraic and r is an integer. Hence the well-known [Ei74],[SS78]:

Theorem 1: *Let L be a regular language. Then its ordinary generating function is a rational function. Hence l_n is asymptotically a finite sum of terms of the form:*

$$C\alpha^n n^r \tag{2}$$

where C, α are algebraic and r is a non-negative integer.

Observe also that any "regular" parameter of a regular language (like the counting of occurrences of letters, subwords etc...) will lead similarly to rational generating functions for which an expansion of the type (2) will hold. Thus for instance the expected number of occurrences of a fixed letter in a random word of length n will be expressible as a quotient of two expansions of the form (2) and hence it cannot be of the form $D\sqrt{n}$.

Context-free Languages.

The family of context-free languages may be defined in the same manner as regular languages, save that now *recursive definitions* are allowed. We are interested in the subfamily of *unambiguous context-free languages*, the ones for which these recursive specifications are unambiguous, each word in the language being constructible in a unique manner.

It is well known from a theorem of Chomsky and Schutzenberger in 1963 that if L is an unambiguous context-free language, then its (ordinary) generating function $l(z)$ is algebraic. This can be checked by arguments very similar to the previous ones, since $l(z)$ will be a component of a rational set of equations and elimination can be performed leading to a unique polynomial equation:

$$P(z, l(z)) = 0 \tag{3}$$

Asymptotic properties of coefficients of algebraic functions are well-known (See *e.g.* [Fla85b]) from a combination of Puiseux expansions _local expansions in fractional powers_ around singularities and a Darboux type of argument. Hence:

Theorem 3: *If L is an unambiguous context-free language, then its generating function $l(z)$ is an algebraic function. Hence l_n is asymptotic to a finite sum of terms of the form:*

$$l^n \sim C\alpha^n n^r \tag{4}$$

where r is a rational number in $\mathbb{Q}/\{-1,-2,-3\cdots\}$ and $C\Gamma(r+1)$ as well as α are algebraic.

Here again for "context-free" parameters, statistical properties of the form $n^{\sqrt{2}}$ or $e^{\sqrt{n}}$ are excluded.

Simple Classes of Trees.

The concept of a *simple class of trees* has been introduced by Meir and Moon in [MM78]. It constitutes a combinatorial analogue of the probabilistic theory of *branching processes* and is a prototype of the form of results we are aiming at: a whole family of classes of combinatorial structures share many common statistical properties concerning for instance height, level of nodes, occurrences of subtrees *etc*... .

Simple classes correspond to recursive specifications that use a variant of the "sequence-of" construction. Specifically, let $\Omega \subset \mathbf{N}$ be a set of integers containing 0. For a given Ω, define the construction $K_\Omega(A) \equiv$ **sequence-of**(A **with** degree$\in\Omega$) as:

$$K_\Omega = \bigcup_{d\in\Omega} A^d$$

where A^d denotes the d-fold cartesian product of A with itself. The simple

class of trees (associated to Ω) is defined by the recursive specification:

$$TREE = \{node\} \times \textbf{sequence-of}(TREE \textbf{ with degree} \in \Omega); \tag{5}$$

thus it is the class of planar trees whose outer degrees of nodes are constrained to be in the set Ω. The corresponding generating function $tree(z)$ satisfies the fixed-point equation:

$$tree(z) = z\,\omega(tree(z)) \quad \text{with} \quad \omega(u) = \sum_{d \in \Omega} u^d \ . \tag{6}$$

Assume also for simplicity of exposition that $\omega(u)$ is a function with the *mixing* property, that is to say it is not of the form $\varphi(u^e)$ for some integer $e \geq 2$.

Meir and Moon argue as follows. Function $t(z) \equiv tree(z)$ is implicitly defined by the equation:

$$f(z, t(z)) = 0 \quad \text{where} \quad f(z,y) = y - z\,\omega(y) \ . \tag{7}$$

By the implicit function theorem, $t(z)$ will be analytic until the smallest positive value ρ such that the system of 2 equations in the unknown (ρ, τ):

$$\begin{cases} f(\rho,\tau)=0 \\ f'_2(\rho,\tau)=0 \end{cases} \tag{8}$$

where $f'_2(x,y) = \dfrac{\partial f(x,y)}{\partial y}$.

From there a bivariate Taylor expansion shows that the dependence between z and t is locally around (ρ, τ) of the form:

$$A(z-\rho) + B(t-\tau)^2 \sim 0$$

or equivalently

$$t(z) \sim c_1 - c_2 \left(1 - \frac{z}{\rho}\right)^{1/2} . \tag{9}$$

The proof then concludes by an appeal to Darboux' theorem by which we can conclude that t_n is equivalent to the coefficient of z^n in the r.h.s., namely:

$$t_n \sim C \rho^{-n} n^{-3/2} \ . \tag{10}$$

Whence:

Theorem 3: *If Ω has the "mixing" property, then the number of trees in the simple class of trees associated to Ω satisfies:*

$$t_n \sim C \rho^{-n} n^{-3/2}$$

As we have announced already, simple classes of trees share many common statistical properties. For instance Flajolet and Odlyzko [FO83] have shown that the expected height of a tree of size n is always of the form $\sim Dn^{1/2}$. In their original paper, Meir and Moon showed general results concerning the profiles of such trees. Steyaert and Flajolet [SF83] showed that the probabilities of occurrence of patterns in such trees have an exponential tail...

The point that is of interest to us here is that despite differences in the definitions, the generating function of trees always has dominant algebraic singularities (branch points of order 1) and similar characterisations will hold true for generating functions of a large number of tree parameters.

3. Random Trains.

Now that our objective is (hopefully) clear, I would like to illustrate some of the phenomena of interest by means of a concrete example, namely a (perhaps somewhat unrealistic) combinatorial model of *trains*. (See Figure 1).

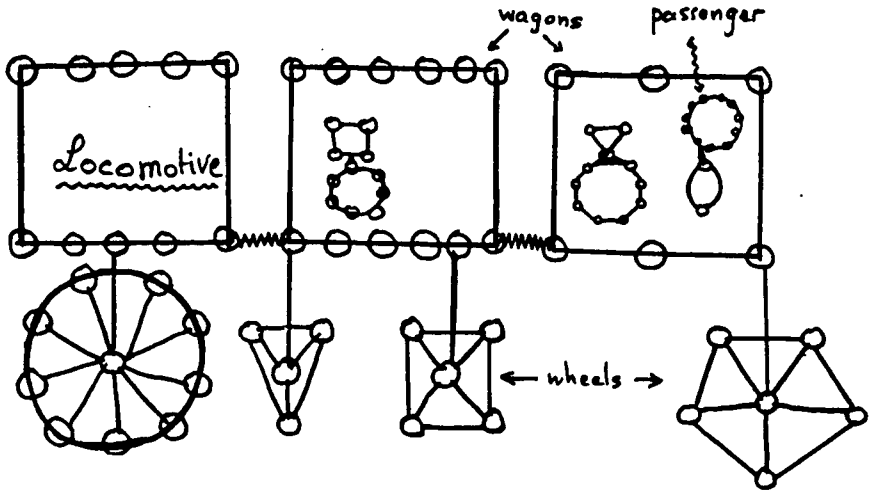

Figure 1: A random train: atoms (nodes) are labelled with distinct and consecutive integers starting from 1 (not shown on the diagram). A train is a complex graph structure in which wheels are defined up to cyclic order and passengers can move freely inside wagons (they are not ordered between themselves inside their respective wagons).

A possible specification for the model is the set of equations (S)

type **(S)**

train = **P–product**(*locomotive*,**P–sequence–of**(*wagon*));

locomotive = **P–sequence–of**(*slice* **with** degree≥1);

slice = **union**(**P–product**(*upper*,*lower*),**P–product**(*upper*,*lower*,*wheel*));

wheel = **P–product**(*center*,**P–cycle–of**(*wheel_element*));

wagon = **P–product**(*locomotive*,**P–set–of**(*passenger*));

passenger = **P–product**(*head*,*belly*);

head,*belly* = **P–cycle–of**(*passenger_element*);

upper,*lower*,*center*,*wheel_element*,*passenger_element* = *atom*(1);

In words: a train is formed of one locomotive and a sequence of wagons; a wagon is like a locomotive but it can contain passengers (considered free to move within their wagon); locomotive and wagons may have wheels that can

rotate freely around their center and they are described as sequences of (vertical) slices to which wheels are possibly attached *etc*...

A train is thus a sort of complex labelled graph structure (although it does not show on Figure 1, distinct integers are associated to different nodes). Natural questions (at least for people travelling by train!) are: What is the expected number of wagons, passengers, wheels in a random train? What is the expected number of (passenger) empty wagons? What is the expected size of the largest wagon? ...

We shall see that many such characteristics of random trains can be analyzed and so, ***en passant***, we shall illustrate how general methods are available to analyze statistical properties of many diverse types of combinatorial structures.

Generating Functions.

Trains are labelled structures, so that from what we saw in the introduction, we should resort to *exponential generating functions* (*e.g.f.*). This is implicit in what follows: $train(z)$ will denote the *e.g.f.* of the class ***train*** (we omit the "hats" for the sake of notational simplicity).

The definitions (S) translate at sight into the system of equations:

$$
\begin{aligned}
&train(z) = locomotive(z).(1-wagon(z))^{-1}; \qquad \text{(T)}\\
&locomotive(z) = slice(z).(1-slice(z))^{-1};\\
&slice(z) = upper(z).lower(z) + upper(z).lower(z).wheel(z);\\
&wheel(z) = center(z).\log(1-wheel\_element(z))^{-1};\\
&wagon(z) = locomotive(z).\exp(passenger(z));\\
&passenger(z) = head(z).belly(z);\\
&head(z),belly(z) = \log(1-passenger\_element(z))^{-1};\\
&upper(z),lower(z),center(z),wheel\_element(z),passenger\_element(z) = z;
\end{aligned}
$$

Now a reasonable *symbolic manipulation system* † starting from these equations will readily provide the result:

† We have been using here the University of Waterloo's MAPLE System, thanks to the courtesy of Gaston Gonnet.

train :=

$$\frac{z^2 + z^3 \ln\left(\frac{1}{1-z}\right)}{\left(1 - z^2 - z^3 \ln\left(\frac{1}{1-z}\right)\right)\left(1 - \frac{z^2 + z^3 \ln\left(\frac{1}{1-z}\right)}{1 - z^2 - z^3 \ln\left(\frac{1}{1-z}\right)}\right) \exp\left(\ln(1-z)^2\right)}$$

It is of some interest to represent the expression of *train*(z) in ***tree form***, in the usual way. What we obtain is depicted in Figure 2. This form also shows the clear connection between subtrees of the expression *train*(z) and subclasses entering the definition of the *train* type.

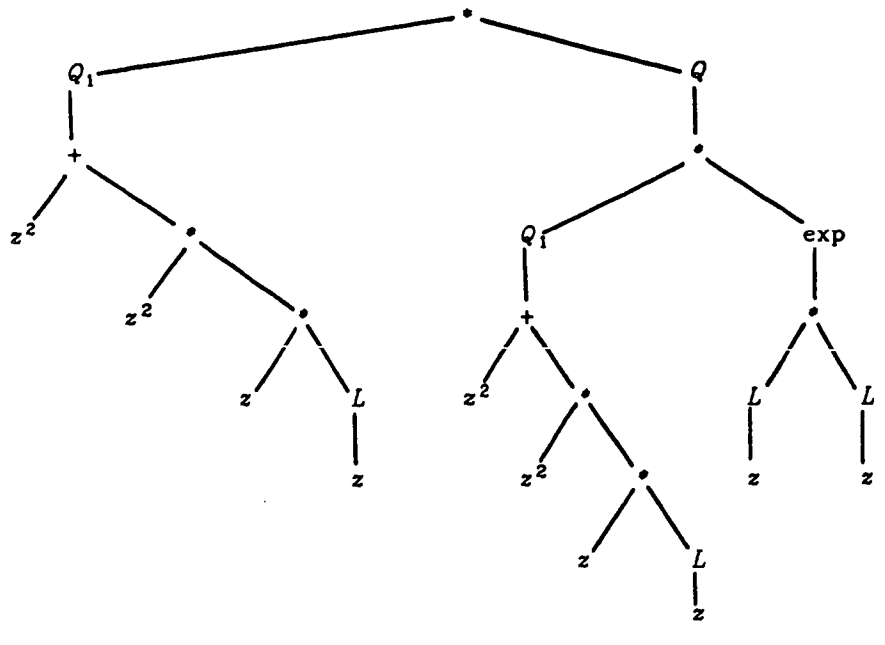

Figure 2: The tree corresponding to function *train*(z). (Here $Q(y)=(1-y)^{-1}$, $Q_1(y)=y.(1-y)^{-1}$ and $L(y)=\log(1-y)^{-1}$.)

Asymptotic Analysis.

We now propose to prove:

Theorem 4: *The number of trains of size n satisfies asymptotically:*

$$train_n \sim C A^n n!$$

where C, A are real constants and A=1.9302981 $\cdots$.

Notice that one can label the nodes of the tree of Figure 2 by the ***radius of convergence*** (*r.o.c.*) of the associated series. That labelling can be constructed inductively starting from the leaves that are monomials labelled with $+\infty$. It is non-increasing along any branch from a leaf to the root. ***Simple rules*** concerning convergence of sums, products, quasi-inverses... might even be used to determine it ***automatically*** as well.

Function $train(z)$ is of the form:

$$train(z) = a(z).\frac{1}{1-b(z)} \tag{11}$$

where $a(z) \equiv locomotive(z)$ and $b(z) \equiv wagon(z)$. Since $b(z)$ which has positive coefficients reaches the value 1 before it becomes singular, we see that $train(z)$ becomes singular at $\rho = b^{(-1)}(1) \equiv 0.5180546 \cdots$ where it has a simple pole and the result of Theorem 4 follows by a singularity analysis with A being equal to ρ^{-1} and:

$$\frac{1}{n!} train_n \sim \frac{a(\rho)}{\rho b'(\rho)} \rho^{-n} = \frac{locomotive(\rho)}{\rho\, wagon'(\rho)} \rho^{-n} . \quad \blacksquare \tag{12}$$

It is important to notice at that stage that the same line of reasoning (11),(12) will generally apply to all classes defined by a *schema* of the form:

$$\mathbf{F = P\text{-}product(A, P\text{-}sequence\text{-}of(B))} \tag{13}$$

or even more simply:

$$\mathbf{F = P\text{-}sequence\text{-}of(B)}$$

provided some mild conditions are satisfied (like **B** has the mixing property; the "density" of **A** measured by the inverse of the r.o.c. of its e.g.f. should be less than that of **B**...).

4. Statistical Properties of Random Trains.

Consider again the *train* type and assume we are interested in the expected number of wagons in a random train of size n. Returning to the form (11) of $train(z)$, we see that the total number of wagons in all trains of size n is:

$$W_n = [\frac{z^n}{n!}]\left\{ \frac{\partial}{\partial u} \frac{a(z)}{1-ub(z)} \Big|_{u=1} \right\} \tag{14}$$

and the e.g.f. $W(z)$ is again a meromorphic function with a ***double pole*** at $z=\rho$. One finds via a singularity analysis that:

$$\frac{1}{n!} W_n \sim \frac{locomotive(\rho)}{\rho^2\, wagon'(\rho)^2} n .$$

The variance of the number of wagons can be studied in the same way; it is

again linear, so that:

Theorem 5: *There exists a non-zero real constant ω_0 such that the expected number of wagons in a random train of size n satisfies:*

$$\overline{W}_n \equiv \frac{W_n}{train_n} \sim \omega_0 n \ .$$

Furthermore the distribution is concentrated: for any ω_1 and ω_2 such that $\omega_1<\omega_0<\omega_2$, there is a vanishing probability that the number of wagons in a random train will be outside that interval: $[\omega_1 n;\omega_2 n]$.

The same type of argument could be applied to the general schema (13-11), again under mild and automatically testable conditions so that Theorem 5 could be extended to a sort of "meta-theorem" concerning structures of the form (13):

$$\mathbf{F = P\text{-}product(A,P\text{-}sequence\text{-}of(B))} \ .$$

In that context, results on the *distribution* of the number of wagons _and more generally of the number of *components* in a "sequence-of" constructs_ can be obtained using saddle point arguments.

Theorem 6: *The probability $\lambda_n^{(k)}$ that a random train of size n has k wagons tends to a* limiting Gaussian distribution.

The proof starts from the generating function expression:

$$\lambda_n^{(k)} = \frac{1}{train_n} [\frac{z^n}{n!}] locomotive(z).wagon(z)^k$$

$$\sim \frac{1}{C}\rho^n \int a(z)b^k(z) \frac{dz}{z^{n+1}} \ .$$

The integral can be evaluated along a circle centered at the origin and passing through the approximate saddle point $\zeta \equiv \zeta(\frac{k}{n})$ defined by:

$$\frac{k}{n}\frac{f'(\zeta)}{\zeta f(\zeta)} - 1 = 0 \ .$$

Studying the dependence of the saddle point approximation as a function of k/n leads to the *local limit theorem*:

$$\lambda_n^{(k)} \to \frac{e^{-\frac{(k-k_0)^2}{2\sigma^2 n}}}{\sigma\sqrt{2\pi n}} \quad \text{with} \quad k_0 = \omega_0 n \ . \quad \blacksquare$$

The random variable "number of wagons" could be called a *critical parameter*: it corresponds to the number of components in a "sequence-of" construct that *determines* the radius of convergence of the global e.g.f. of the structure under consideration. Other *subcritical parameters* counting the number of constructs in the tree below *wagon*, like number of passengers, number of wheels in wagons will be analyzable in the same way:

Theorem 7: *The parameters number of passengers and number of wagon wheels have $O(n)$ expectation and $O(n)$ variance so that their distributions are concentrated around their means.*

The proof here is again achieved by a singularity analysis. With our previous notations, the bivariate generating functions are all of the form:

$$\frac{a(z)}{1-\varphi(z,u\psi(z))} \quad \text{with} \quad \varphi(z,\psi(z)) = b(z)\,. \ \blacksquare \tag{15}$$

In contrast, *non-critical* parameters coming from the **A** component in the top-level product will have radically different behaviours. Consider for instance the parameter L defined as the size of the locomotive in a random train. The cumulated value over all trains of size n, L_n, admits the e.g.f:

$$\frac{\partial}{\partial u}\left\{locomotive(uz)\frac{1}{1-wagon(z)}\right\}\Big|_{u=1} \tag{16}$$

a series that now has only a simple pole at $z=\rho$. Whence:

Theorem 8: *The expected size of the locomotive in a random train of size n has asymptotically finite mean and finite variance. The expectation satisfies:*

$$\bar{L}_n \equiv \frac{L_n}{train_n} \sim locomotive'(\rho)\,.$$

Furthermore, the distribution of the "locomotive size" parameter admits a limiting distribution *that has an exponential tail.*

Regarding the distribution result, we see that the probability $\pi_n^{(k)}$ that the locomotive has size k in a random train of size n tends to the limiting distribution:

$$\pi_n^{(k)} \sim \frac{wagon_k \frac{\rho^k}{k!}}{wagon(\rho)}\,,$$

which does have an exponential tail from our assumption: r.o.c.(*wagon*)>r.o.c.(*train*). ■

Similar results will hold true for other non-critical parameters like the number of slices, wheels... of the locomotive. All will have asymptotically finite mean and variance and an exponential tail.

Let us also give a brief indication on the filling of passengers in wagons. The generating function for the total number of wagons containing k passengers, k a fixed integer, is of the form:

$$p^{(k)}(z) = \frac{\partial}{\partial u}\left\{\frac{a(z)}{1-c(z)(e^{d(z)}+(u-1)\frac{d^k(z)}{k!}}\right\}\Big|_{u=1} \tag{17}$$

whence, after a singularity analysis which shows that the corresponding probabilities are proportional to

$$\frac{d^k(\rho)}{k!}$$

the result:

Theorem 9: *The probability that a random wagon in a random train of size n has exactly k passengers tends to the* limiting Poisson distribution

$$e^{-\lambda}\frac{\lambda^k}{k!} \quad \text{with} \quad \lambda = passenger(\rho)\ .$$

An only slightly modified argument applied now to functions of the form:

$$\frac{a(z)}{1-f(z)(\frac{1}{1-g(z)}+(u-1)g^k(z))} \tag{18}$$

recording cases of application of a subcritical "sequence-of" construct _like the number of slices in wagons_ leads to:

Theorem 10: ***The probability that a random wagon in a random train has exactly k slices tends to the*** limiting geometric distribution

$$\frac{\mu^{k-1}}{1-\mu} \quad \text{with} \quad \mu = slice(\rho)$$

Other interesting problems leading to more complicated equations are *maximal* parameters like size of fattest passenger, largest wagon, most crowded wagon... but we do not have time to discuss them here. For instance, Feller and Knuth have discussed the expected length of the longest run of ones in a random binary string; Shepp and LLoyd study the size of the largest cycle in a permutation, Flajolet and Odlyzko consider the size of the largest component in a random functional graph. Results about the largest component in a random allocation have been obtained by Gonnet [Go81]...

We summarise in Table 1 some of the main characteristics of random trains.

Param.	Exp/Var.	Distribution
loco. size	$<O(1),O(1)>$	L.D., exp. tail
loco. wheels	$<O(1),O(1)>$	L.D., exp. tail
loco. slices	$<O(1),O(1)>$	L.D., exp. tail
numb. wagons	$<O(n),O(n)>$	L.D.: Gaussian
numb. passeng.	$<O(n),O(n)>$	concentr.
numb. wag. wheels	$<O(n),O(n)>$	concentr.
weight passeng.	$<O(n),O(n)>$	concentr.
size random wagon	$<O(1),O(1)>$	L.D.
numb. pass. rand. wagon	$<O(1),O(1)>$	L.D.: Poisson
numb. slices rand. wagon	$<O(1),O(1)>$	L.D.: Geometric

Table 1: A summary of characteristics of random trains. ("L.D." means existence of a limiting distribution; "exp. tail" means exponential tail; "concentr." means that the distribution is concentrated in the sense of Theorem 5).

5. Extensions.

There is not much that is special about random trains. Only the situation was made a little easier since the *critical construct* that determines the radius of convergence of the whole series is a **P–sequence–of** which induces a *pole*. However, other critical constructs can also be analysed in similar general terms.

Assume for instance we consider a *railways system* defined by:

$$\textbf{type } railways\_system = \textbf{P–set–of}(train)\,. \tag{19}$$

Then (19) translates into:

$$railways\_system(z) = \exp(train(z))\,. \tag{20}$$

If looking for the expected number of trains in a random railways system (There are quite a few in real life!) of size n, we find it to be a quotient of 2 Taylor coefficients:

$$\bar{T}_n = \frac{[z^n]\frac{\partial}{\partial u}\exp(\frac{a(z)}{1-b(z)})\Big|_{u=1}}{[z^n]\exp(\frac{a(z)}{1-b(z)})}\,. \tag{21}$$

Quantity (21) is also the quotient of 2 integrals

$$\frac{\int\exp(\frac{a(z)}{1-b(z)})\frac{a(z)}{1-b(z)}\frac{dz}{z^{n+1}}}{\int\exp(\frac{a(z)}{1-b(z)})\frac{dz}{z^{n+1}}} \tag{22}$$

which may be both evaluated using an approximate saddle point of the form:

$$\zeta = \zeta(n) = \rho - K\sqrt{n}$$

from which follows:

Theorem 11: *The expected number of trains in a random railways system of size n is asymptotic to $K_1\sqrt{n}$.*

The work of Hayman on "admissible functions" ("admissible" has here a meaning different from that of Section 1) enters as a necessary ingredient in these analysis: Hayman gives general conditions under which saddle point methods can be applied to Cauchy integrals.

Another important and interesting question is to incorporate *recursive definitions* in the specification language for combinatorial structures. Some of the methods introduced by Darboux, Polya, Meir, Moon and others should also be of help there.

6. General Conclusions.

This paper contains many (partly unsupported) claims that I will now summarise:

1. Most of current analyses of algorithms and combinatorial enumerations deal with objects having relatively short specifications, say about 20 lines of PASCAL programme for instance. One could use "off-the-shelf" technology developed for these analyses in order to analyze many structurally complex objects without a considerable increase in the mathematical machinery. In that enterprise, symbolic manipulation systems (MACSYMA, MAPLE ...) are of course a necessity.

2. Many mathematically interesting phenomena should appear when considering various statistics of these structures; furthermore many of them should be *simple*. We have hinted at a few such phenomena when discussing linear *versus* constant expectation (in relation with critical/subcritical/non-critical parameters), concentration of distributions (or variance estimates), exponential tails, Gaussian, Poisson or Geometric limiting distributions *etc*.... From an analytical standpoint, the problem is largely the study of some simple analytic functionals operating on series with positive coefficients.

3. Concerning the analysis of algorithms, two directions are conceivable:

- Abstract Analysis of Algorithms: by this I mean the study of simple algorithms using *decomposition primitives* that are dual of the combinatorial constructs we have examined. For instance, what can be said of a schema of the form:

```
type F = P-sequence-of(B)
procedure search(b : B);
var c : component_of B;
    begin
    c := first(b);
    while not Π(c) do
        c = next(c);
    end;
```

for various predicates Π,under various cost measures associated to the testing Π and to moving to the next component? In plain words, examine problems like: How much time does it take to find a (passenger) empty wagon? ...

- Specialise this type of study to more restricted data types like expression trees, heap-ordered trees and the like, taking into account more specific mechanisms (See *e.g.* [FS82], [St84] for such an attempt regarding expression trees).

As a final word, it seems that there might be a promising area of research in the investigation of statistical properties of *combinatorial schemata* and the closely related domain of the average case analysis of *programme schemata* which were extensively studied for their structural properties in the seventies.

References:

[DB57] N. G. DE Bruijn: *Asymptotic Methods in Analysis*, North Holland P.C. (1957); reprinted by Dover (1981).

[Ei74] S. Eilenberg: *Automata, Languages and machines*, Academic Prees, New-York (1974).

[Fla85a] P. Flajolet: Mathematical Analysis of Algorithms and Data Structures, in *A Graduate Course in Computation Theory*, Computer Science Press, (1985, to appear).

[Fla85b] P. Flajolet: Ambiguity and Transcendence, in *Proc. I.C.A.L.P.*, Nafplion (July 1985). To appear in *Lecture Notes in Comp. Science*.

[FO83] P. Flajolet and A. Odlyzko: The Average Height of Binary Trees and Other Simple Trees, *J. of Computer and System Sc.* **25** (1983), pp. 345-369.

[FS82] P. Flajolet and J-M. Steyaert: A Complexity Calculus for Classes of Recursive Search Programs over Tree Structures, in *Proc., 22nd IEEE Sump. on Found. of Comp. Theory (FOCS)*, Nashville (1982), pp. 386-393.

[Go81] G. Gonnet: Expected Length of the Longest Probe Sequence in Hashing, *J.A.C.M.* **28** (1981), pp. 289-304.

[GJ83] I. Goulden and D. Jackson: *Combinatorial Enumerations*, J. Wiley (1983).

[He77] P. Henrici: *Applied Computational and Complex Analysis*, J. Wiley, New-York (1977).

[MM78] A. Meir and J. W. Moon: On the Altitude of Nodes in Random Trees, *Canad. J. of Math.* **30** (1978), pp. 997-1015.

[O83] A. Odlyzko: Periodic Oscillations of Coefficients of Power Series That Satisfy Functional Equations, *Adv. in Math* **44** (1982), pp. 180-205.

[SF83] J-M. Steyaert and P. Flajolet: Patterns and Pattern-Matching in Trees: An Analysis, *Inf. and Control* **58** (1983), pp. 19-58.

[SS78] A. Salomaa and M. Soittola: *Automata-Theoretic Aspects of Formal Power Series*, Springer Verlag (1978).

[St84] J-M. Steyaert: *Complexite et Structure des Algorithmes*, Thesis, University of Paris VII (1984), 215p.

Imprimé en France

par

l'Institut National de Recherche en Informatique et en Automatique